

SALTcymru_

Document 1

Overview report on relevant open source software and standards in speech and language technology, and their possible enhancement for use in Wales in a pre-competitive research stage

**Prepared by the
Language Technologies Unit (Canolfan Bedwyr), Bangor University**

April 2008



Llywodraeth Cynulliad Cymru
Welsh Assembly Government

This document was prepared as part of the SALT Cymru project, funded by the Welsh Assembly Government under the Knowledge Exploitation Fund's Knowledge Exchange Programme, reference HE 06 KEP 1002

Five software packages are evaluated as part of SALT Cymru, covering key areas of SALT as pinpointed by the project survey and users' and developers' feedback. They are as follows:

- Festival: speech synthesis
- Sphinx: speech recognition
- UIMA: the semantic web (intelligent web searching; keyword and trend spotting from text)
- Tesseract: optical character recognition
- Moses: machine translation

This document contains an overview of all the packages that have been evaluated, summarising the findings on each one, and highlighting priorities for future development.

Document 2 offers an overview of Festival, a popular open-source framework designed for speech synthesis development. It shows that its development for the Welsh context is at an intermediate level, offering functional voices that are well-integrated with popular operating systems. However, the voice quality for Welsh is currently lower than that for British or American English.

Document 3 discusses Sphinx, a set of open-source packages for speech recognition. In contrast to speech synthesis, speech recognition has not been significantly developed for Welsh, and any project will start from a very low level of development. However, this issue is to some extent being dealt with by a project carried out by the Language Technologies Unit at Bangor University, and funded by the Welsh Language Board for the financial year 2008/09, to develop basic speech synthesis for Welsh. While this will not result in anything approaching the level of maturity for English, it represents a foundation upon which future development can be built.

Document 4 investigates the UIMA framework, which offers a structured platform for information extraction. This provides significant possibilities for future development and high potential worth to Welsh SMEs, as it facilitates applications such as intelligent web searching, information retrieval and extraction. Unstructured information is the largest and the fastest growing source of information to business and organisations. Welsh SMEs could therefore be both developers and users of such technology. Any applications that could structure such information in a multilingual environment would provide savings and efficiencies for businesses beyond the original developers.

Document 5 discusses the Tesseract OCR (optical character recognition) system, which is available as a free, open-source standalone application for English, but due to its open nature, allows other languages to be included within its framework. There is currently no OCR technology that produces satisfactory results when scanning Welsh and bilingual Welsh/English printed text. The ability to accurately digitize Welsh language texts would benefit many sectors and expertise developed in the process could be put to commercial use in other languages.

Document 6 discusses the Moses machine translation system, an open framework for developing the automatic translation by computer of one human language to another. It presents the results of a small pilot study in developing machine translation for the Welsh-English language pair. The results are highly accurate for language close to the domain for which the system was originally trained, i.e. proceedings and legislation from the Welsh Assembly Government. It is less accurate for language divergent from this. This does, however, show promise for development of machine translation for Welsh and English and its use in certain restricted contexts, as long as sufficient bilingual textual data can be found for use in development.

In addition to the resources and potential resources for Welsh mentioned in the previous sections, there are other general aids to speech and language technology that have a role in future development in the field. Examples include corpora of speech and language, required for applications such as speech recognition and machine translation. Some work has already been undertaken in these fields. The SpeechDat Welsh database, collected between 1996-99, presented digitized recordings of 2000 Welsh speakers over telephone lines. It was designed for the development of voice-driven telephony services. A written electronic corpus, CEG, consists of a million words of Welsh, and has been used widely, particularly in the development of spelling and grammar checkers for the language, and in speech synthesis applications. Welsh still lags behind more widely-spoken languages in language resources of the kind, and there is a clear need both for a larger written corpus and a non-telephone speech corpus. Both of these are seen as essential aids for the development of SALT in any language.

It is noted that all the packages described in these documents are open-source ones. This is a deliberate decision. Such packages can be developed in a non-state aid environment such as the ones that exist for many funding streams exploited by Welsh academia and industry. A successful model for this, used by the Language Technologies Unit, Bangor University, in previous projects, has been to develop tools under a liberal free software license. This license allows the tools to be used in other projects, whether commercial or free, without restriction. Therefore, the tools can be developed through grant aid and versions of them released to the public at no cost. Outside the scope of these projects, other versions can be developed from the free ones, either by academia under exclusive or non-exclusive contracts to businesses, or by businesses possessing sufficient knowledge to develop the products themselves. Such non-free products can be tailored to businesses' individual needs, or can provide refinements not present in the free versions. Businesses can then resell or further develop these versions, offering them to end-users at cost.